University of Toronto Scarborough
Department of Computer and Mathematical Sciences
STAD29 / STA 1007 (K. Butler), Final Exam
April 24, 2018, 2:00–5:00pm (3 hours)

Aids allowed:

- My lecture overheads (slides)
- Any notes that you have taken in this course
- Your marked assignments
- Non-programmable, non-communicating calculator

Before you begin, complete the signature sheet, but sign it only when the invigilator collects it. The signature sheet shows that you were present at the exam.

This exam has 12 numbered pages of questions. Please check to see that you have all the pages.

In addition, you should have an additional booklet of output to refer to during the exam. Contact an invigilator if you do not have this.

Answer each question in the space provided (under the question). If you need more space, use the backs of the pages, but be sure to draw the marker's attention to where the rest of the answer may be found.

The maximum marks available for each part of each question are shown next to the question part. In addition, the total marks available for each page are shown at the bottom of the page, and in the table on the next page.

The University of Toronto's Code of Behaviour on Academic Matters applies to all University of Toronto Scarborough students. The Code prohibits all forms of academic dishonesty including, but not limited to, cheating, plagiarism, and the use of unauthorized aids. Students violating the Code may be subject to penalties up to and including suspension or expulsion from the University.

Last name: _____

First name: _____

Student number: _____

For marker's use only:

Page	Points	Score
1	9	
2	12	
3	7	
4	6	
5	6	
6	8	
7	9	
8	11	
9	8	
10	9	
11	10	
12	9	
Total:	104	

You may assume throughout this exam that the code shown in Figure 1 of the booklet of code and output has already been run.

1. An experiment is done to test the effect of a toxic substance on insects. At each of six dose levels, 250 insects are exposed to the substance and the number of insects that die is counted. We are interested in whether there is an association between dose and how likely an insect is to die, and if so, what kind of relationship it is.

   The data are shown in Figure 2 (in the booklet of code and output). The values are separated by spaces, not tabs.

   (a) (2 marks) Explain briefly why logistic regression is a sensible way to assess our problem of interest.

   (b) (1 mark) Give R code to read in the data, as shown in Figure 2, from a file called `toxicity.txt` in the current folder.

   (c) (4 marks) Give R code to fit a suitable logistic regression for these data, predicting the probability of death, and to display the results.

   (d) (2 marks) Part of the output from your logistic regression is shown in Figure 3. Is there evidence that the probability of death depends on dose, and does the probability of death go up or down as dose increases? Explain briefly.

2. When children are learning to solve a puzzle, does it make any difference how much encouragement they get? In a study, 20 children were randomly assigned to one of four treatment groups:

   - Constant reward
   - Frequent reward
   - Infrequent reward
   - No reward

   according to how often they received a reward or encouragement from the experimenter as they were trying to solve a puzzle. For each child, the number of attempts required to solve the puzzle was recorded (a smaller number of attempts means that the child solved the puzzle more quickly).

   The researchers have some specific research hypotheses in mind:

   - Any reward (average of) will produce faster solution than no reward
   - Constant reward will produce faster solution than (the average of) frequent and infrequent reward.
   - Frequent reward will produce faster solution than infrequent reward.

   A "faster solution" is one that takes *fewer* attempts. The data are shown in Figure 4.

   (a) (2 marks) Explain briefly why the researchers decided to use contrasts rather than the standard ANOVA F-test followed by Tukey.

   (b) (3 marks) I created an ordered factor from `reward` in Figure 5. Write contrasts C1, C2, C3 that will enable the researchers to test their research hypotheses, in the order that they are given above. (This is most easily done as three lines of R code.)

   (c) (3 marks) What R code (one or two lines, depending how you write it) would arrange it so that the `lm` in Figure 6 does the right tests (instead of the default use of `Constant` as a baseline and comparing everything with that)?

   (d) (4 marks) What do you conclude from the analysis in Figure 6, in the context of the data? Note: there is a subtle feature of the analysis which, for full marks, you should address.

3. Diabetes is a disease in which a person has high blood sugar over a long period. Two ways of treating diabetes are a special diet, and regular injections of insulin. The aim of these treatments is to keep blood sugar at an acceptable level.

   A study was carried out in which 20 adults wth diabetes were randomly assigned to one of two treatment groups, `diet` and `insulin`. The study lasted six months. Each subject had their fasting blood sugar measured at the beginning and at the end, and the change in fasting blood sugar level, `FBS_change`, was recorded. It is also suspected that the fasting blood sugar levels, and therefore their change, will depend on the age of the subject. (Blood sugar levels also depend on what the subject has eaten, and so are measured when the subject has *not* eaten for a specified number of hours.)

   The dataset is shown in Figure 7.

   (a) (3 marks) A plot is shown in Figure 31. (This is at the end of the Code and Output because it is in colour.) What does this plot tell you about any possible relationships between change in fasting blood sugar and: (i) treatment, (ii) age of subject, (iii) the treatment-age combination? Explain briefly.

   (b) (2 marks) Why is an analysis of covariance suitable here? Explain briefly.

   (c) (2 marks) An analysis of covariance is shown in Figure 8. What do you conclude from it, in the context of the data?

(d) (2 marks) A further analysis is shown in Figure 9. Is it necessary to look at this? If it is necessary, what do you conclude from it, in the context of the data? If it is not necessary, explain briefly why not.

(e) (2 marks) More output from the analysis of the previous part is shown in Figure 10. Is it helpful to look at this? If it is helpful, what do you conclude from it, in the context of the data? If it is not, explain briefly why not.

(f) (2 marks) Compare your conclusions from Figure 31 and from parts (c) through (e). Describe briefly how they are consistent or inconsistent.

4. What might influence a person's ability to remember a list of words? One possibility is that if the words on the list are related in some way, it would be more difficult to remember which specific words were on the list. For example, if the list contained "baseball", "soccer" and "hockey", you might have trouble remembering whether the list also contained "football" or "basketball", but if there was only one sport, you might more easily remember which one it was.

   A study was conducted in which subjects were presented with three different lists of words at three different times. The three word lists contained the same number of words and the same mixture of common and rare words, but varied according to whether, and how, the words were related to one another. Specifically, subjects had to learn word lists of these different types:

   - `unrelated`: the words are unrelated to each other.
   - `semantic`: the words have similar meanings (to other words on the list).
   - `phonological`: the words have similar sounds (to other words on the list).

   The order in which a subject was given the word lists was randomized.

   The response variable was the number of words on each list that the subject successfully remembered. The data are shown in Figure 11. (There are sixteen subjects; the subjects with IDs 10 and 11 withdrew from the study.)

   (a) (2 marks) Why do you think that each subject was given the word lists in a random order?

   (b) (2 marks) What is it about the design of this experiment that makes a repeated measures analysis necessary?

   (c) (2 marks) Some analysis is shown in Figure 12. In the second line of code (the one with `lm` in it), why did I need the number `1` after the squiggle? Explain briefly.

Question 4 continues… This page: _____ of possible 6 points.

(d) (3 marks) What do you conclude from the analysis in Figure 12, in the context of the data?

(e) (1 mark) A spaghetti plot is shown in Figure 13. Some of the points on the plot are joined by lines. Which ones?

(f) (2 marks) In the code above Figure 13, I used `gather`. Why did I need to do that? I am looking for an answer specific to this task, not a general description of what `gather` does.

(g) (2 marks) Compare what you conclude from Figure 12 and Figure 13. Do the conclusions seem to be consistent or inconsistent? Explain briefly.

5. A rootstock is part of a plant, often a root, from which new above-ground growth can be produced. This can be done by "grafting" one or more cuttings from another plant onto the rootstock in such a way that they will grow into healthy plants. In the data shown in Figure 14, six apple tree rootstocks are used. On each of these, eight apple trees are grown. The individual trees are labelled in the column called `row`; there are 48 of them. The idea is that trees grown on the same rootstock should have consistent properties, and different properties than those grown on other rootstocks. To see whether that is true for these apple trees, four variables are measured for each tree:

   - `girth4`: the girth (circumference) of the tree trunk after 4 years (mm, multiplied by 100)
   - `extension`: extension growth at 4 years (metres)
   - `girth15`: the girth of the tree trunk after 15 years (same units as the girth after 4 years)
   - `weight` of above-ground portion of tree after 15 years (in pounds, multiplied by 1000).

   The purpose of multiplying by the different numbers was to get values that were comparable in size.

   Our aim is to see whether the trees from the different rootstocks differ on any of the measured variables, and, if so, which ones.

   (a) (2 marks) Explain briefly how MANOVA will help us address at least part of our aim (and which part it will help with).

   (b) (3 marks) Give code to run the MANOVA and to display the results. Give some thought to whether the `rootstock` variable is the right kind of thing. (You may assume that you already have a data frame `rootstocks` as in Figure 14.)

   (c) (2 marks) How do you know that the discriminant analysis in Figure 16 was worth doing? Explain briefly.

   (d) (2 marks) Looking at Figure 16, how many linear discriminants do you think you should pay attention to? Explain briefly.

Question 5 continues… This page: _____ of possible 9 points.

(e) (2 marks) What would make a tree have a very *negative* score on `LD1`? Explain briefly.

(f) (2 marks) What would make a tree have a large positive score on `LD2`? Explain briefly.

(g) (3 marks) Look at Figure 17. Find an observation (a tree) that is misclassified. Which row number is it in the data frame? Was this tree's rootstock close to being gotten correct? Explain briefly, showing that you know what the columns of the data frame represent.

(h) (2 marks) Look at Figure 33. This figure is in colour, at the end of the Code and Output. The overall misclassification rate is around 50%. Is that about what you would expect, looking at this plot? Explain briefly.

(i) (2 marks) In Figure 33, find a tree that scores *low* (ie., very negatively) on LD1. Find that tree in Figure 14. Are its variable values high and/or low as you predicted for such a tree in part (e)? Explain briefly.

6. A chemical analysis was carried out of 178 wines from a certain region in Italy. Thirteen different quantitative variables were measured for each wine. These variables are listed in Figure 18. Some of the data is shown in Figure 19. (Each wine has a text ID beginning with V, since the Italian for "wine" is "vino".)

   (a) (2 marks) Why is it not possible to run discriminant analysis on these data? Explain briefly.

   (b) (2 marks) Why would a K-means cluster analysis be more appropriate for these data in their current form than a hierarchical cluster analysis?

   (c) (2 marks) A calculation is shown in Figure 20. Explain briefly what the calculation does and why it is necessary.

   (d) (2 marks) Explain briefly why it is necessary to look at a scree plot before running the K-means analysis.

(e) (4 marks) Suppose you have a function called `ss` that accepts as input a number of clusters and a data frame, and returns as output the total within-cluster sum of squares for a K-means cluster analysis with that many clusters on the input data frame.

Give code to obtain a data frame containing the total within-cluster sum of squares for each number of clusters from 2 through 20 inclusive, using the appropriate data set, and use that data frame to make a scree plot.

(f) (2 marks) A scree plot is shown in Figure 21. What do you conclude from this plot? Explain briefly.

(g) (3 marks) I obtained a 3-cluster solution from K-means. (This may or may not be a good number of clusters.) To make a graph of this solution, I ran a discriminant analysis, using the three clusters as known groups. I also drew a biplot. This, and the code to produce it, are shown in Figure 32.

According to the biplot, what do you think is the most important way in which my clusters 1 and 3 differ? Explain briefly.

7. Five people were trying on ski boots in a store on a Friday evening in January. They were each asked about factors which might influence which ski resort they would go to. The questions on the questionnaire were designed to assess:

   - `cost`: cost of ski ticket
   - `lift`: speed of ski lift
   - `depth`: depth of snow
   - `powder`: moisture of snow (drier snow, called "powder", is better for skiing on).

   Some people might attach greater importance to some of these variables, and some might attach greater importance to others.

   After the questionnaires were completed, a score on each of the above variables was computed for each person, with a higher score indicating greater importance on that variable for that person. The data are shown in Figure 22. Our aim is to see whether these four variables can be summarized by fewer variables.

   (a) (2 marks) A principal components analysis is done in Figure 23 and a scree plot is obtained in Figure 24. What do you conclude from the scree plot? Explain briefly.

   (b) (2 marks) In Figure 25, which two variables are the most important part of each relevant component? Explain briefly.

   (c) (2 marks) A biplot is shown in Figure 26. How does this support your answer to the previous part? Explain briefly.

   (d) (2 marks) What do the variables in the first principal component have in common with each other? What do the variables in the second principal component have in common with each other?

   (e) (2 marks) The biplot in Figure 26 suggests that factor analysis will produce a clearer result than principal components. How? Explain briefly.

8. 1681 residents of twelve areas of Copenhagen (in Denmark) were classified according to four categorical variables:

   - `Sat`: Satisfaction with their housing; low, medium or high
   - `Infl`: Feeling of influence on building management; low, medium or high
   - `Type`: Type of housing; tower, apartment, atrium or terrace
   - `Cont`: Degree of contact with neighbours, low or high

   There is also a column `Freq` showing how many residents fell into that combination of categories.

   The data set (some) is shown in Figure 27. The data set reached me as a `data.frame` so I used `as_tibble` to display it nicely.

   Our aim is to discover which of the other variables are associated with satisfaction, and if so, how they are related.

   (a) (3 marks) Some analysis is shown in Figure 28. There is more than I showed (which is omitted). Describe the general process by which I got from `housing.1` to `housing.5`, and why I stopped at `housing.5`.

   (b) (1 mark) Explain very briefly why I am not interested in the `Infl:Cont` and `Type:Cont` terms in Figure 28, even though they are significant.

   (c) (2 marks) Use Figure 29 to describe the relationship between the amount of contact with neighbours and a resident's satisfaction with their housing.

   (d) (3 marks) Use Figure 30 to explain what the significant `Sat:Infl:Type` term tells you about the data.